SCHOOL OF STATISTICS
UNIVERSITY OF THE PHILIPPINES DILIMAN

WORKING PAPER SERIES

Predictive Accuracy of Fitted Logistic Regression Model Using Ranked Set Samples

by

Kevin Carl P. Santos
and
Erniel B. Barrios

UPSS Working Paper No. 2012-02
January 2012

School of Statistics
Ramon Magsaysay Avenue
U.P. Diliman, Quezon City
Telefax: 928-08-81
Email: updstat@yahoo.com

Predictive Accuracy of Fitted Logistic Regression Model Using Ranked Set Samples

Kevin Carl P. Santos
Erniel B. Barrios
School of Statistics
University of the Philippines Diliman

Abstract

Separation of likelihood and rare events are two interrelated problems in fitting the logistic regression model. We propose to address these issues by drawing the sample using ranked set sampling. An extensive simulation study was conducted to assess the performance of a logistic regression model fitted from ranked set samples and compared to those estimates using simple random samples. RSS performs best in small populations regardless of the distribution of the binary response variable in the population. As the sample and population sizes increase, the predictive ability under RSS also improves but it stabilizes to become comparable to SRS. Furthermore, RSS mitigates the problem of separation of likelihood especially when the population size is relatively large. In addition, RSS can be an alternative sampling scheme to Inverse Sampling in obtaining samples involving rare characteristics without necessarily blowing up the sample size. RSS provides sample into the estimation of logistic regression models high predictive accuracy and keeps costs at low levels.

Keywords: logistic regression model, ranked set sampling, rare characteristics, separation of likelihood

I. Introduction

Logistic regression model has become a very popular statistical technique in analyzing categorical variables. Aside from its interpretability and less stringer assumptions, it also provides tools in selection of the important predictors or independent variables. In epidemiology, logistic regression model can be used to determine the risk factors of a disease (Steinmann, et al., 2007). In business analytics, it can be used to find out if a client would churn or not (Owczarczuka, 2010). It can also be used to predict the outcome of a licensure examination (Wanvarie and Sathapatayavongs, 2007) using demographic profiles and academic records of the examinees as covariates.

There are natural constraints in the implementation of logistic regression. Menard (1995) enumerated some of issues like specification error, multicollinearity, and numerical problems like zero cells and complete separation, that usually limit the implementation of logistic regression. Note that the first two problems are also common in linear regression. Specification error occurs when the functional form of the explanatory variables is not correctly specified, or that some important predictors are left out, or no data is available, or that it cannot be measured.

Furthermore, inclusion of irrelevant independent variables in the model would increase the standard errors of the parameter estimates, resulting to unreliable inference, see Menard (1995).

More often than not, the independent variables are naturally correlated. Multicollinearity should be avoided in building models because it would bring problems in estimation and in hypothesis testing.

Zero cells and separation of likelihood often threaten the numerical implementation of maximum likelihood estimation in logistic regression. Zero cell count may lead to some odds ratio estimates equal to infinity or zero. Moreover, because of the zero cells, the iteration for fitting the logit model does not converge or the standard errors of the estimates are very large, see Agresti (1996).

In modelling binary response variables, the separation of likelihood, or simply called separation is a common problem. This occurs when one or more of a model's covariates perfectly predict some binary outcome as noted by Zorn (2005). Albert and Anderson (1984) as cited by Lesaffre and Albert (1989) compared complete and quasi-complete separation, the latter denoting the case when such perfect prediction occurs only for a subset of observations in the data. Zorn (2005) further observed that when separation of likelihood occurs, the likelihood function becomes monotone resulting to infinite maximum likelihood coefficient estimates for the predictor/s which cause/s the separation.

Zorn (2005) identified the implications of complete separation. First, if there is a covariate which perfectly predicts the response variable, then there is no variation in the dependent variable that is needed to be explained by the other independent variables. Therefore, the corresponding coefficient estimates for other predictors will be zero. Second, since the likelihood is monotone or flat, this will yield large or infinite standard error estimates. For the case of quasi-complete separation, the coefficient estimate for the variable which causes the separation and its standard errors will be infinite, but the other covariates in the model may not be affected.

There are several solutions for the separation problem in the literature. The most popular is dropping the separating variable(s) in the model. Clogg et al. (1991) suggests adding "artificial" data across the different patterns of (categorical) covariates and analysis is done on the modified data. Another approach is to use exact logistic regression that allows estimation of the coefficients even in the presence of empty cells and complete separation. The exact logistic regression uses the method of conditional maximum likelihood in performing exact inference for a parameter yielding exact p-values rather than approximations. This method prevents infinite estimated odds ratios or confidence intervals with one side equal to infinity, see Agresti (1996). However, Zorn (2005) pointed out that exact logistic regression may result to degenerate estimates when relatively sparse data or small number of observations in particular patterns of categorical predictors is present.

Another approach in solving the problem of separation of likelihood is the modified score procedure, Firth (1993). The procedure modifies the maximum likelihood estimation by penalizing the score equation. The method reduces the bias on the maximum likelihood estimates

for the coefficients of the logistic regression model. Moreover, the procedure does not produce infinite coefficient estimates. Heinze and Schemper (2002) noted the advantages of the modified score procedure over some known approaches to solve monotone likelihood problem.

Clogg et. al (1991) considered the possibility of resolving the separation problem by adding "artificial" data across the different patterns of the categorical predictors and then conducting the analysis in the resulting data in the usual manner. This paper aims to consider sampling strategy as a possible solution to the separation problem. Since the separation problem usually arises from the existence of "patterns" among the data on the predictors, then it is possible that the problem is avoided if the likelihood of such "pattern" is minimized. As noted by Menard (1995), the problem of separation of likelihood can be attributed to the data at hand. Hence, this study explores the possibility of using Ranked Set Sampling (RSS) as a remedy to the problem of monotone likelihood or separation.

2. Logistic Regression with Rare Events

Rare events and unbalanced distribution of the two categories ("event" and "non-event") in the population leads to some constraint in the estimation of logistic regression model, its predictive ability is also affected. Prediction of rare events using logistic regression, gathering data and estimation procedure would also be very crucial.

King and Zeng (2001) observed that the problem with rare events is at least of two types. The first one is that logistic regression can sharply underestimate the probabilities of rare events. The second type in analyzing data with rare events is the inefficiency of the data collection method used. Some corrections were recommended and showed that a prior correction in logistic regression is needed if the functional form and explanatory variables seem appropriate; otherwise, the authors' corrected version of weighting with rare event corrections would likely perform better.

Maalouf and Trafalis (2011) recommended the use of robust weighted kernel logistic regression in case of rare events. The estimation procedure combines the rare events corrections to logistic regression with truncated newton methods and are applied to kernel logistic regression (KLR). Maalouf and Trafalis (2011) noted that according to the significance test, rare event weighted kernel logistic regression (RE-WKLR) is more accurate than both support vector machines and truncated newton method in kernel logistic regression. Therefore, using the RE-WKLR leads to greater accuracy in predicting rare events.

Furthermore, Maalouf and Trafalis (2011) reviewed known sampling methods dealing with rare events and reported that two of the basic sampling strategies are under-sampling, which eliminates units from the majority class or category, and over-sampling, which adds more units from the minority class or category. King and Zeng (2001) recommended the use of under-sampling of the majority class or category when logistic regression will be utilized. However, this sampling scheme induces biased in the estimates of the coefficient.

3. Ranked Set Sampling

In order to make reliable inferences while keeping costs at low-level, McIntyre (1952) proposed the ranked set sampling procedure. The sampling procedure offered to raise precision with availability of visual assessment of ranking units and without measuring the actual variable of interest. Stoke (1977) proposed the use of auxiliary information in the implementation of ranked set sampling. The idea is similar to sampling with probability proportional to size (PPS), the auxiliary variable or a frugal measurement is used in ranking the individuals or units. Stoke (1977) concluded that the level of precision depends on the degree of the relationship of the auxiliary and the target variables. It is assumed that the frugal measurement or concomitant variable is cheap and/or very easy to obtain. In medical studies, quantitative genetics, and ecological and environmental studies, some attributes can be easily obtained or quantified; however, some variables of interest are oftentimes time-consuming and/or expensive or impossible to measure. The use of auxiliary variables broadened the applications of the ranked set sampling scheme.

Mapping is usually done in the study area to collect information on the concomitant variable before the target response variable and the covariates are measured. In addition, the frugal measurement, similar to the auxiliary variable for sampling with probability proportional to size (PPS), should be correlated to the target variable. The choice of the concomitant variable is very crucial and should be based on some sound theory. The sampling procedure of balanced RSS is as follows:

1. Consider a simple random sample (SRS) of size *k*, where *k* is called the set size, from the population of size N. The SRS of size *k* is obtained from the sampling frame only, it is not necessary to physically draw the sampled units or individuals. The sampling frame should contain an auxiliary variable that is correlated with the target variable (or, equivalently, some frugal measurement can be easily collected). We use the concomitant variable to rank the units in the SRS. The smallest or the 1^{st} order statistic will be the 1^{st} element in the sample of size *k*. After obtaining the 1^{st} order statistic, the *k-1* elements will be disregarded, and are eligible for subsequent selection.

2. Obtain another SRS of size *k* and rank these units according to the concomitant variable. The 2^{nd} order statistic or 2^{nd} smallest unit will be the 2^{nd} element in the sample of size *k*. Again, the *k-1* units will be discarded and are made available for subsequent draws. Repeat the process until the k^{th} order statistics is selected to complete the sample of size k.

3. Steps 1and 2 generates one cycle of samples. Repeat the cycle *m* times to obtain a sample of size $n = mk$.

The literature on ranked set sampling provide estimation procedure of parameters such as the mean (Takahasi and Wakimoto, 1968), proportion (Chen, et al., 2005) and variance (Stokes, 1980) We explore in this paper the use of RSS in model-building.

Muttlak (1995) suggested the use of RSS in estimating the parameters of the simple regression model. The objective was to increase the precision of the estimators of the slope and

intercept relative to SRS design. It was reported that balanced RSS is advantageous over SRS in providing more efficient estimates of regression coefficients.

Twidwell (2000) fitted a model which relates the fish length, as a proxy variable for the length exposure, and the tissue concentration of mercury. The aim of the study was to determine the area of east Texas affected by contamination of mercury. Since the process involved in the study was expensive and destructive, RSS was implemented to minimize cost and increase precision. A simulation study by Twidwell (2000) showed that a balanced ranked set sample would lead to a slight improvement as compared to a simple random sample in regression analysis.

Murff and Sager (2006) investigated the application of balanced RSS to Ordinary Least Squares (OLS) regression analysis through analytical and simulation studies. Some important findings are: if the ranking is done on the independent variable (RSSX), equivalently and often more efficient slope and intercept estimators may be achieved by performing SRS of three more sample items or units; if ranking is done on the dependent variable (RSSY), the gains in efficiency in equivalent sample size obtained by the slope estimator appears at best comparable to that of the RSSX slope estimator and at worst less than a SRS slope estimator of the same sample size.

Chen et al. (2004) considered the dependent variable Y and the concomitant variables $\underline{\mathbf{X}}$ in the classical linear regression model $Y = \beta_0 + \underline{\beta}' \underline{X} + \varepsilon$ where $\underline{\beta}$ is the vector of the coefficients of the independent or concomitant variables and ε is the stochastic error term, which is assumed to be normally distributed with mean 0 and variance σ^2.

The estimators can be obtained using least squares estimation procedure. Let

$$\overline{\underline{X}}_{RSS} = \frac{1}{mk}\sum_{r=1}^{k}\sum_{i=1}^{m}\underline{X}_{[r]i} \qquad \overline{Y}_{RSS} = \frac{1}{mk}\sum_{r=1}^{k}\sum_{i=1}^{m}Y_{[r]i}$$

$$\underline{X}_{RSS} = (\underline{X}_{[1]1},\ldots,\underline{X}_{[1]m},\ldots,\underline{X}_{[k]1},\ldots,\underline{X}_{[k]m})'$$

$$\underline{Y}_{RSS} = (y_{[1]1},\ldots,y_{[1]m},\ldots,y_{[k]1},\ldots,y_{[k]m})'$$

where *k* is the set size and *m* is the number of cycles. The least squares estimators of β_0 and $\underline{\beta}$ are, respectively, given by:

$$\hat{\beta}_0 = \overline{Y}_{RSS} - \underline{\hat{\beta}}'_{RSS}\,\overline{\underline{X}}_{RSS}$$

$$\underline{\hat{\beta}}'_{RSS} = [\underline{X}'_{RSS}(\underline{I} - \frac{\underline{1}\underline{1}'}{mk})\underline{X}_{RSS}]^{-1}\underline{X}'_{RSS}(\underline{I} - \frac{\underline{1}\underline{1}'}{mk})\underline{Y}_{RSS}$$

These estimators are unbiased and are at least asymptotically as good as their counterparts based on an SRS. Estimating the regression coefficient, balanced RSS and SRS are asymptotically equivalent. Similar to the finding of Murff and Sager (2006), Chen et al. (2004) concluded that balanced RSS cannot do much for the improvement for the estimation of the regression coefficients. If this is the main concern, unbalanced RSS should be considered.

Özdemir and Alptekinesin (2007) investigated the parameter estimation in multiple linear regression models using RSS as well. They examined the variances of the estimators using a

Monte Carlo simulation study. RSS estimators of the coefficients of the regression model are more efficient than that of SRS when the sample size is small.

Chen and Wang (2004) used RSS in regression analysis in a lung cancer study. The objective was to investigate the relationship of smoking status and three biomarkers: polyphenol DNA adducts, micronuclei, and sister chromatic exchanges. In this case, measuring the biomarkers is very expensive but the determination of the smoking level of individuals can be easily determined. Hence, RSS is appropriate because this sampling procedure promises low costs but relatively high efficiency compared to SRS. The set size k=10 in RSS yield remarkable improvement of the optimal sampling schemes over SRS.

We evaluate the use of RSS in estimating a qualitative response model/discrete choice model.

4. Simulation Studies

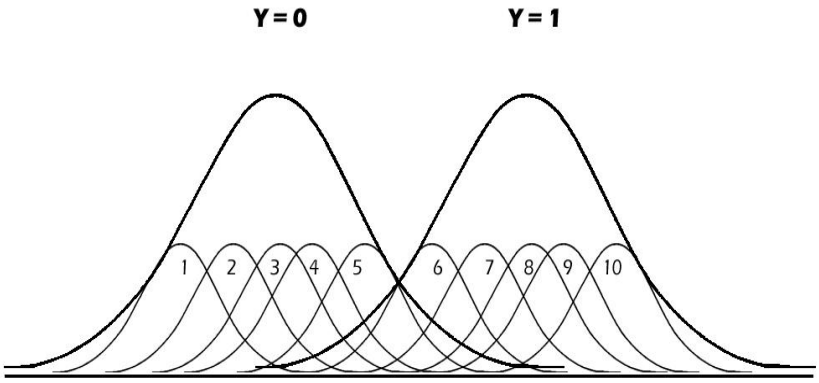
A simulation study is designed with goal of determining the predictive accuracy of the logistic regression model using samples obtained using RSS and compare the results to those using simple random sampling (SRS). We consider four predictors that included the concomitant variable.

Assume that the binary response variable is expensive to measure and the readily available concomitant variable used in the ranking process is a viable predictor in the model. The idea in choosing the concomitant variable is that its high/low values are associated with Y=1 or "success" while low/high values are associated with Y=0 or "failure".

The percentage of correct classification (PCC), sensitivity, and specificity are used as measures of predictive accuracy. The average number of separation of likelihoods in estimating the parameters of the logistic regression model is also recorded. This will help determine if high PCC, specificity, or sensitivity are brought about by separation and to verify the claim if RSS is indeed a possible solution to the problem of monotinicity of likelihoods.

Sampling through RSS could ensure that the population will be well-represented because low values of the concomitant variable that correspond to "failures" would be obtained in the sample as well as high values which correspond to "success" as shown in Figure 1. This is also the reason why RSS work well for small populations which are usually heterogeneous. The sample gathered is evenly distributed to different parts of the distribution of the target variable. This idea should also work in the case of rare events or unbalanced distribution of success and failure because RSS will generate "enough" representation of the binary variable. The resulting samples could mitigate the likelihood problem thereby producing reasonable parameter estimates and, subsequently, better predictive ability of the estimated logistic regression model.

Figure 1. Distribution of the response variable Y



The population size (N=1,000, 5,000 and 10,000) is varied to determine how RSS will affect the sample and later in the estimation of the logistic regression model. RSS usually performs well in small populations while SRS generally yield representative samples in large populations of when the population is homogeneous. Furthermore, sampling rate (n=1%, 3%, 5% of N) is varied to assess the efficiency of estimates of the parameters of the model to varying sample sizes. Moreover, we also investigate the ability of RSS in resolving the separation of likelihood problem in small population with small sample size.

Different distributional assumptions of the concomitant variable are considered such as normal and Poisson. A count ranking variable contains less information than a continuous variable. The mean and variance were allowed to vary. High degree of variation implies that the concomitant variable contain more information on the target variable Y, while lower variance means less information on Y.

The likelihood separation problem would likely occur when the binary distribution of the response variable is severely unbalanced. Thus, different levels of unbalancedness would help in the assessment of the effect of sample selection (RSS) in estimating the logistic regression model. The distribution of the "success" and "failure" categories in the population is varied as follows: balanced (50%-50%), moderately balanced (40%-60%), moderately unbalanced (25%-75%), and severely unbalanced (10%-90%), for the "success"-"failure" proportions.

We fix the set size in RSS at k=10. In Figure 1, RSS would obtain samples from each of the small bell-shaped curves representing the distribution of the 1st until the 10th order statistics. Hence, RSS would most likely generate "enough" representative of the population.

Table 1 shows the specifications in the implementation of RSS in obtaining the samples that will be used later on in the estimation of the logistic regression model. Since the sample size $n = mk$ then the number of cycles is $m = \frac{n}{10}$.

Table 1. Number of Cycles in RSS by Sample Size and Population Size

Population Size (N)	Sample Size (n)	Number of Cycles (m)
1000	10	1
	50	5
	100	10
5000	50	5
	250	25
	500	50
10000	100	10
	500	50
	1000	100

5. Results and Discussion

In Table 2, the percent of correctly classified observations (PCC) of the fitted logistic regression model using ranked set samples (RSS), are generally higher than that of using simple random samples (SRS). The largest difference, around 30%, can be noticed for n = 10 and N = 1,000. But, as the sample size increases, the difference of the PCC's of the two sampling scheme depreciates, and is true across all population sizes. Similar observation can be noticed for the specificity and sensitivity of RSS and SRS. Furthermore, even if only 1% of the population is sampled, RSS will less likely result to separation of likelihood. The highest occurrence of separation of likelihood is 16.5% for the extreme case (n=10, N=1,000). In SRS, however, there is a very high chance of having a separation of likelihood especially for small populations. For moderate and large populations, separation of likelihood happens more often only in very small sample size, but increases in sample size lowers the chance of separation of likelihood. Compared to RSS, SRS always encounter a very large number of times of monotonicity of likelihood. Very high values of PCC's for SRS are actually consequence of separation of likelihood.

Table 3 shows that, even under moderately balanced distribution of the binary variable, RSS performs better than SRS. The difference of the evaluation measures between the two sampling schemes is largest for the small populations (N=1,000).

Sensitivity is the percentage of observations which are predicted to belong in the "success" category given that they actually belong there. In Table 3, sensitivity under different sample sizes and population sizes for RSS are generally larger than those in SRS. However, when the population is getting large, the sensitivity of the models estimated using samples from two sampling designs becomes comparable. It can also be noticed that RSS had separations of likelihood only in the extreme scenario (small population, small sample size). This suggests that it is very unlikely for RSS to encounter the problem of separation for small populations, given that the sample size is 5-10% of the population size, and for average to large populations

regardless of the sample size. Compared to RSS, SRS has a very high chance of occurrence of separation of likelihood, specifically when the sample size is only 1%. Generally, the results are almost similar to that of the balanced case. This means that when the distribution of the binary variable is moderately balanced, the predictive accuracy of the fitted logistic regression model is not far from the balanced case.

The difference between the PCC's for RSS and SRS are largest when the population is small (N=1,000) for moderately unbalanced case as shown in Table 4. As the population size increases, the PCC's become comparable for the two sampling designs. In addition, the specificity for RSS is higher than for SRS when the population size is small. As the sample size increases, the specificities also increase. RSS has greater advantage over SRS when the population size is small. As the population size increases, the specificity for the two sampling designs becomes comparable. It can also be observed that when the population is moderate (5,000) or large (10,000), it is less likely for RSS to cause a separation of likelihood. Hence, even if only 1% of the population size is taken as the sample, separation would not be a problem anymore in RSS, this is not the case in SRS.

Sensitivity for the severely unbalanced case is the lowest since only 10% of the population is classified as "success" category. Hence, it is more difficult in this case to generate samples coming from the "success" category.

Table 5 summarizes the results when the distribution of Y is severely unbalanced in the population. Even if the distribution of Y is severely unbalanced, separation of likelihood is still very unlikely to happen in RSS most especially when 5-10% of a small population is sampled or at least 1% for a medium- to large-sized population. In SRS, however, there is a very high change of obtaining a sample with very small number of observations belonging in the "success" category. Based on the sensitivity in Table 5, RSS still has greater advantage over SRS when the population size is small, i.e. N=1,000. But, as the sample size and population size increase, the proportion of correct classification improves for SRS. The sensitivity for large population (10,000) for RSS and SRS are comparable. Large differences for sensitivity of RSS and SRS can be noticed in cases where the sample size is 1%. This suggests that if it is very expensive to obtain a sample, gathering only 1% of the population using RSS would suffice and would yield better predictive accuracy over SRS. This further suggests that RSS provides an option to inverse sampling when obtaining samples of units with rare characteristic while maintaining a fixed sample size.

Table 2 - Measures of Predictive Accuracy for Balanced Case by Sample Size and Population Size

Sample Size	50-50%	N=1,000						% of Separation	
		RSS			SRS				
		PCC	Specificity	Sensitivity	PCC	Specificity	Sensitivity	RSS	SRS
10		96.65	96.58	96.68	63.06	60.08	61.14	16.5	100
50		97.95	97.51	98.4	93.14	93.52	92.5	1	99.25
100		97.85	97.36	98.32	96.08	96.21	95.33	0.25	95.5
Sample Size	50-50%	N=5,000						% of Separation	
		RSS			SRS				
		PCC	Specificity	Sensitivity	PCC	Specificity	Sensitivity	RSS	SRS
50		97.46	96.89	98.04	93.53	93.59	93.24	1.5	99
250		97.09	96.91	98.72	97.27	97.22	97.29	0	41.25
500		97.8	96.83	98.71	97.52	97.4	97.61	0	9.75
Sample Size	50-50%	N=10,000						% of Separation	
		RSS			SRS				
		PCC	Specificity	Sensitivity	PCC	Specificity	Sensitivity	RSS	SRS
100		97.64	96.79	98.49	96.09	95.98	96.15	2.75	92.25
500		97.74	96.52	98.93	97.38	96.86	97.87	0	6.75
1000		97.63	96.34	98.9	97.5	96.77	98.21	0	2.5

Table 3 – Measures of Predictive Accuracy for Moderately Balanced Case by Sample Size and Population Size

Sample Size	60-40%	N=1,000						% of Separation	
		RSS			SRS				
		PCC	Specificity	Sensitivity	PCC	Specificity	Sensitivity	RSS	SRS
10		97.53	98.06	96.63	62.68	63.32	56.04	62	100
50		98.20	98.85	97.15	93.25	94.25	91.53	0	100
100		98.10	98.78	97.03	96.30	96.95	95.28	0	100
Sample Size	60-40%	N=5,000						% of Separation	
		RSS			SRS				
		PCC	Specificity	Sensitivity	PCC	Specificity	Sensitivity	RSS	SRS
50		98.10	98.40	95.58	93.50	94.58	91.63	0	100
250		98.18	98.65	97.45	97.70	98.18	96.90	0	50
500		98.23	98.58	97.63	97.90	98.38	97.10	0	0
Sample Size	60-40%	N=10,000						% of Separation	
		RSS			SRS				
		PCC	Specificity	Sensitivity	PCC	Specificity	Sensitivity	RSS	SRS
100		98.18	98.55	97.58	96.50	97.05	95.53	0	100
500		98.25	98.68	97.58	97.90	98.33	97.18	0	0
1000		98.15	98.60	97.48	98.05	98.50	97.35	0	0

Table 4 - Measures of Predictive Accuracy for Moderately Unbalanced Case by Sample Size and Population Size

Sample Size	75-25%	N=1,000						% of Separation	
		RSS			SRS				
		PCC	Specificity	Sensitivity	PCC	Specificity	Sensitivity	RSS	SRS
10		96.98	97.98	93.62	60.97	65.79	41.04	22	100
50		98.40	99.08	96.48	93.75	95.48	87.75	0	100
100		98.43	99.18	96.10	96.50	97.6	93.08	0	100
Sample Size	75-25%	N=5,000						% of Separation	
		RSS			SRS				
		PCC	Specificity	Sensitivity	PCC	Specificity	Sensitivity	RSS	SRS
50		98.01	98.69	95.93	93.49	95.47	86.87	0.8	99.8
250		98.28	98.78	96.75	97.78	98.53	95.53	0	50
500		98.38	98.80	97.20	98.15	98.78	96.35	0	0
Sample Size	75-25%	N=10,000						% of Separation	
		RSS			SRS				
		PCC	Specificity	Sensitivity	PCC	Specificity	Sensitivity	RSS	SRS
100		98.28	98.75	96.88	96.35	97.55	92.60	0	100
500		98.38	98.85	98.98	98.20	98.85	96.23	0	0
1000		98.38	98.83	97.05	98.30	98.88	96.63	0	0

Table 5 - Measures of Predictive Accuracy for Severely Unbalanced Case by Sample Size and Population Size

Sample Size	90-10%	N=1,000						% of Separation	
		RSS			SRS				
		PCC	Specificity	Sensitivity	PCC	Specificity	Sensitivity	RSS	SRS
10		95.53	97.91	72.76	59.40	65.09	17.20	23.8	99.5
50		96.38	98.50	78.23	92.45	95.43	61.13	0	75
100		96.90	98.78	79.50	94.95	97.45	70.88	0	75
Sample Size	90-10%	N=5,000						% of Separation	
		RSS			SRS				
		PCC	Specificity	Sensitivity	PCC	Specificity	Sensitivity	RSS	SRS
50		96.83	98.63	81.00	92.63	95.60	61.38	0	75
250		96.93	98.85	79.58	96.48	98.48	77.88	0	75
500		96.90	98.80	79.45	96.78	98.70	79.28	0	0
Sample Size	90-10%	N=10,000						% of Separation	
		RSS			SRS				
		PCC	Specificity	Sensitivity	PCC	Specificity	Sensitivity	RSS	SRS
100		96.90	98.93	78.83	95.30	97.70	70.98	0	75
500		97.05	98.93	79.63	96.85	98.73	79.63	0	25
1000		97.05	99.00	79.40	96.95	98.83	79.75	0	0

6. Illustration

The proposed methodology is applied to farming households. From the Census of Agriculture in 2002 conducted by the National Statistics Office (NSO), very few farmers plant lettuce, cauliflower, and asparagus. These are some of the high value crops that grow only in few areas in the Philippines. Instead of using the information per farmer, the variables were aggregated in the *barangay* level. Thus, the units of measurement considered in this illustration are the *barangays*.

The target binary variable in this case is defined as follows:

$$Y_i = \begin{cases} 1, & \text{if } i^{th} \textit{ barangay} \text{ has at least one farmer who plants either lettuce, cauliflower, or asparagus} \\ 0, & \text{otherwise} \end{cases}$$

Only 2.18% (273 out of 12519) of all the *barangays* in the country have farmers planting the rare high value crops. Therefore, the distribution of the target variable Y is severely unbalanced. This case is much more unbalanced than the case 90%-10% considered in the simulation scenarios. The objective here is to predict if rare high value crops are grown in a *barangay* or not using the estiamted logistic regression model.

The concomitant variable to be used in ranking the *barangays* is the proportion of farmers in a *barangay* who have post-graduate degrees. Growing these crops requires more advanced agronomic concepts; hence, it is expected that as farmers achieve more advanced degrees, they will likely to be knowledgeable in growing these crops. This indicator is indeed strongly correlated to the target binary variable. In fact, the odds ratio between the two variables is approximately 10. Since the data set considered is a census, the concomitant variable is readily available. The predictors to be used in the estimation procedure are mean age of the farmers in the *barangay*, total farm area, total area harvested from 2001-2002, proportion of irrigated farms, and proportion of farmers who have post-graduate degrees.

The predictive accuracy of the fitted model using RSS and SRS are given in Table 6.

Table 6 - Percentage of Correctly Classified Predicted *Barangays* which Plant Rare High Value Crops by Sample Size

Sample Size	RSS				SRS			
	PCC	Sensitivity	Specificity	Separation	PCC	Sensitivity	Specificity	Separation
140	93.2	22.2	98.4	Yes	96.1	20	99.2	No
660	98.4	16.7	100	Yes	98.5	0	99.8	No
1000	97.8	14.3	99.8	No	97.3	0	99.7	No
1350	97.9	24.1	99.7	No	98.3	0	99.9	No

The primary interest here is to predict correctly if a *barangay* has at least one farmer planting the rare high value crops. The sensitivity for RSS are higher than those of SRS. This suggests that using RSS in fitting the logistic regression model has greater advantage in predicting a *barangay* with farmers planting lettuce, asparagus, or cauliflower correctly over SRS. This result supports the findings from the simulation scenarios because the concomitant variable and the binary response variable have a strong relationship. Thus, the predictions of the estimated logit model are more accurate when RSS is used instead of SRS.

7. Conclusion

Samples drawn using RSS yield better estimates of the logistic regression model compared to SRS. However, as the population and sample sizes increase, the two sampling designs become comparable in terms of percentage of correctly classified observations in the estimated logistic regression model. The predictive ability of the estimated logistic model using RSS performs best when the population size is small. This is true for different proportions of the "success"-"failure" in Y in the population. The predictive ability of the fitted model using both RSS and SRS usually decreases as the distribution of Y in the population becomes severely unbalanced.

If the costs in obtaining samples, most especially of rare events or characteristics, are too high, small sampling rate would be sufficient in estimating a logistic regression model provided that RSS is used. RSS can be an alternative sampling scheme to inverse sampling in obtaining rare events or characteristics without blowing up the sample size. Furthermore, the correlation between the concomitant variable and the binary response variable matters more than the nature of the frugal measure.

The use of RSS in drawing of samples to be used in fitting a logistic regression model can prevent the problem of separation of likelihood even when the distribution of Y is severely unbalanced. For moderately-sized sampling rate in small populations, or with a very small sampling rate in large populations, using RSS in drawing samples to be used in estimating the logit model would unlikely encounter the separation of likelihood problem.

Acknowledgment

The authors would like to thank the Statistical Research and Training Center (SRTC) for providing funding support for this study.

References

Agresti, A. (2002). *Categorical Data Analysis*. 2nd edition. Hoboken, NJ: John Wiley & Sons, Inc.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. NY: John Wiley & Sons, Inc.

Albert, A., and Anderson, J.A. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika,* 71 (1), 1-10.

Chen, H., Stasny, E. and Wolfe, D. (2005). Ranked Set Sampling for Efficient Estimation of a Population Proportion. *Statistics in Medicine*, 24 (21), 3319-3329.

Chen Z., and Wang, Y. (2004). Efficient Regression Analysis with ranked-set sampling. *Biometrics*, 60 (4), 997-1004.

Chen, Z., Bai, Z., and Sinha, B. (2003). *Ranked Set Sampling: Theory and Application*. Springer.

Clogg, Clifford C., Rubin D. B., Schenker, N., Schultz, B. and Weidman, L. (1991). Multiple Imputation of Industry and Occupation Codes in Census Public–Use Samples Using Bayesian Logistic Regression. *Journal of the American Statistical Association,* 86, 68–78.

Firth, D. (1993). Bias Reduction of Maximum Likelihood Estimates. *Biometrika,* 80 (1), 27–38.

Heinze, G. and Schemper, M. (2003). A Solution to the Problem of Separation in Logistic Regression. *Statistics in Medicine*, 21 (21), 2409–2419.

King, G., and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis,* 9, 137-163.

Lesaffre, E., and Albert, A. (1989). Partial Separation in Logistic Discrimination. *Journal of the Royal Statistical Society*, Series B 51,109–116.

Maalouf, M., and Trafalis, T. (2011). Robust weighted kernel logistic regression in imbalanced and rare events data. *Computational Statistics and Data Analysis*, 55,168-183.

McIntyre, G.A. (1952). A method of unbiased selective sampling using ranked sets. *Australian Journal of Agricultural Research*, 3, 385–390.

Menard, S. (1995). *Applied Logistic Regression Analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-106. Thousand Oaks, CA: Sage.

Muttlak, H. (1995). Parameters Estimation in a Simple Linear Regression Using Rank Set Sampling. *Biometrical Journal,* 37 (7), 799-810.

Murff, E., and Sager, T. (2006). The relative efficiency of ranked set sampling in ordinary least squares regression. *Environmental and Ecological Statistics*, 13, 41-51.

Owczarczuka, M. (2010). Churn Models for Prepaid Customers in the Cellular Telecommuncation Industry using Large Data Marts. *Expert Systems with Applications,* 37 (6), 4710-4712.

Özdemir, Y. and Alptekinesin, A. (2007). Parameter Estimation in Multiple Linear Regression Models using Ranked Set Sampling, Commun.*Fac.Sci.Univ.Ank.Series A1,* 56 (1), 7-20.

Steinmann, P., et al. (2007). Helminth infections and Risk Factors Analysis among residents in Eryuan country, Yunnan province, China. *Acta Tropica,* 104, 38-51.

Stokes, S. (1977). Ranked set sampling with concomitant variables. *Communications in Statistics - Theory and Methods,* A6 (12), 1207–1211.

Takahasi, K. and Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics,* 20, 1–31.

Twidwell, S (2000). Bioaccumulation of Mercury in Selected East Texas Water Bodies. Surface Water Quality Monitoring Team, AS-180, Texas Natural Resource Conservation Commission, Austin TX.

Wanvarie, S. and Sathapatayavongs, B. (2007). Logistic Regression Analysis to Predict Medical Licensing Examination of Thailand. *Annals of Academy of Medicine Singapore,* 36, 770-773.

Zorn, C. (2005). A Solution to Separation in Binary Response Models. *Political Analysis*, 13 (2), 157-170.